1. Attention exploration

(a)

i. 对于序列中每一个值都有一个 α（α 在 0 到 1 之间），并且 $\sum \alpha_i = 1$.

ii $k_j^T q \gg k_i^T q$，$\forall i \neq j$

iii $\alpha_j \approx 1$ $c = \sum_i v_i \alpha_i = v_j \alpha_j = v_j$

iv 单词的key与query点积越大，则该词注意力权重越高，得到的注意力输出则~~越~~接近词的value

(b) $v_a = c_1 a_1 + c_2 a_2 + \cdots + c_m a_m = Ac$

$v_b = d_1 b_1 + d_2 b_2 + \cdots + d_q b_q = Bd$

构造M

$Ms = v_a \Rightarrow M(v_a + v_b) = v_a$

$\Rightarrow Mv_a = v_a \quad Mv_b = 0$

$Mv_a = MAc = c \quad Mv_b = MBd = 0$

由正交性

$a_j^T a_i = 0 \quad j \neq i$

$a_j^T a_j = 1 \quad j = i$

$a_j^T b_k = 0 \quad \forall j, k$

$\therefore$ 令 $M = A^T$

$M v_a = M A c = A^T A c = c$

$M v_b = M B d = A^T B d = 0$

满足定式，故 $M = A^T$

(ii) $c \approx \frac{1}{2}(v_a + v_b)$

$\therefore \alpha_a = \alpha_b = \frac{1}{2}$

$\Rightarrow k_a^T q = k_b^T q \gg k_i^T q \quad \forall i \neq a, b$

设 $q = \beta(k_a + k_b) \quad \beta \gg 0$

则 $k_a^T q = \beta \quad k_b^T q = \beta \quad k_i^T q = 0 \ (i \neq a, b)$

$$\alpha_a = \alpha_b = \frac{\exp(\beta)}{n-2+2\exp(\beta)} \approx \frac{\exp(\beta)}{2\exp(\beta)} = \frac{1}{2}$$

(c)

i. 由于协方差矩阵很小，

故 $k_i \approx \mu_i$

由(b) i 得

$q = \beta(\mu_a+\mu_b) \qquad \beta \gg 0$

ii) $\mu_i^T\mu_i = 1$

$\Sigma_a = \alpha I + \frac{1}{2}(\mu_a\mu_a^T)$

$k_a \approx \gamma\mu_a \qquad \gamma \sim N(1, 0.5)$

$k_i \approx \mu_i \qquad i \neq a$

$k_a^T q \approx \gamma\mu_a^T\beta(\mu_a+\mu_b) \approx \gamma\beta$

$k_b^T q \approx \mu_b^T\beta(\mu_a+\mu_b) \approx \beta$

$$\therefore \alpha_a \approx \frac{\exp(\gamma\beta)}{\exp(\beta)+\exp(\gamma\beta)} = \frac{1}{1+\exp(\beta(1-\gamma))}$$

$$\alpha_b \approx \frac{\exp(\beta)}{\exp(\beta)+\exp(\gamma\beta)} = \frac{1}{1+\exp(\beta(\gamma-1))}$$

$\gamma \sim N(1, 0.5)$

$\gamma \to 0.5 \quad \alpha_a \approx \frac{1}{1+\infty} \approx 0 \qquad \alpha_b \approx \frac{1}{1+0} \approx 1$

$\gamma \to 1.5 \quad \alpha_a \approx \frac{1}{1+0} = 1 \qquad \alpha_b \approx \frac{1}{1+\infty} = 0$

$C = \alpha_a v_a + \alpha_b v_b$

$\gamma \to 0.5 \quad C \simeq v_b$

$\gamma \to 1.5 \quad C \simeq v_a$

(d)

i 设 $q_1 = \beta \mu_a \quad q_2 = \beta \mu_b \quad \beta >> 0$

则 $C_1 \simeq v_a \quad C_2 \simeq v_b$

$\therefore C = \frac{1}{2}(C_1 + C_2) \approx \frac{1}{2}(v_a + v_b)$

ii $q_1 = \beta \mu_a \quad q_2 = \beta \mu_b$

则 $k_a^T q_1 = \gamma \mu_a^T \beta \mu_a = \gamma \beta$

$k_b^T q_2 = \mu_b^T \beta \mu_b = \beta$

$\alpha_a \simeq 1 \quad \alpha_b \simeq 1$

$C_1 \approx v_a \quad C_2 \doteq v_b$

$\therefore C \simeq \frac{1}{2}(v_a + v_b)$